FiD-Ex: Improving Sequence-to-Sequence Models for Extractive Rationale Generation

Kushal Lakhotia[†*] **Bhargavi Paranjape**[‡] **Asish Ghoshal**[†] **Wen-tau Yih**[†]
Yashar Mehdad[†] **Srinivasan Iyer**[†*]
[†]Facebook AI [‡]University of Washington
{kushall, aghoshal, scottyih, mehdad, sviyer}@fb.com
bparan@cs.washington.edu

Abstract

Natural language (NL) explanations of model predictions are gaining popularity as a means to understand and verify decisions made by large black-box pre-trained models, for tasks such as Question Answering (QA) and Fact Verification. Recently, pre-trained sequence to sequence (seq2seq) models have proven to be very effective in jointly making predictions, as well as generating NL explanations. However, these models have many shortcomings; they can fabricate explanations even for incorrect predictions, they are difficult to adapt to long input documents, and their training requires a large amount of labeled data. In this paper, we develop FiD-Ex[1], which addresses these shortcomings for seq2seq models by: 1) introducing sentence markers to eliminate explanation fabrication by encouraging extractive generation, 2) using the fusion-in-decoder architecture to handle long input contexts, and 3) intermediate fine-tuning on re-structured open domain QA datasets to improve few-shot performance. FiD-Ex significantly improves over prior work in terms of explanation metrics and task accuracy on five tasks from the ERASER explainability benchmark in both fully supervised and few-shot settings.

1 Introduction

While large pre-trained language models (Devlin et al., 2019; Raffel et al., 2019; Lewis et al., 2020) with hundreds of millions of parameters have made super-human performance possible on various NLP datasets, they lack transparency into their decision making process, which can adversely affect user trust in their predictions. Recent works have proposed the use of natural language (NL) rationales (Lei et al., 2016; DeYoung et al., 2020; Latcinnik and Berant, 2020) as a means to either obtain an understanding of the reasoning process of models, or as a human-readable snippet for users to verify predictions (Lipton, 2018). Figure 1 presents examples of extractive textual rationales for two QA tasks from the ERASER benchmark (DeYoung et al., 2020)[2]. Recently, Narang et al. (2020) show that sequence to sequence (seq2seq) models outperform previous methods at generating textual rationales for various explainability benchmarks. However, seq2seq models can fabricate rationales even for wrong predictions, are hard to scale to datasets involving several, long evidence documents, and, require large amounts of expensive rationale annotated data for training. In this paper, we introduce FiD-Ex, to alleviate these problems and enhance seq2seq models to achieve significant gains in rationale generation performance.

Camburu et al. (2020) find that models that generate free-form NL explanations can tailor them to convincingly justify incorrect model predictions,

Q: Is Sanskrit the first language of the world ?

The early Jain scholar ... of Sanskrit. Sanskrit belongs to the Indo - European family of languages. It is one of the three ancient documented languages that likely arose from a common root language now referred to as ...

Answer: False

Q: Where does Frodo live ?
Choices: Tunnels, Underground, Somewhere nearby

... Tasha oohed in awe. I said, "Frodo's been visiting you, eh ?" Malaquez said, "Your pet ?" "Hardly. He lives around here somewhere. I suppose he was attracted to the commotion up the hill." ...

Answer: Somewhere nearby

Figure 1: Example questions, answers and corresponding passages from the BoolQ and MultiRC datasets from the ERASER benchmark (DeYoung et al., 2020). Annotated rationales are highlighted. Note that rationales can be multi-sentence and non-contiguous.

[*]Equal Contribution.
[1]github.com/facebookresearch/fidex

[2]In this work, we use textual rationales and NL explanations interchangeably.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3712–3727
November 7–11, 2021.

for example, generating "There is no dog in the image" to justify an *no* prediction on the image of a dog. Although recent seq2seq models (Narang et al., 2020) obtain state of the art performance on rationale generation benchmarks, they are vulnerable to having similar behaviours and can hallucinate new facts by tapping into stored world knowledge in the language model parameters. In order to retain their effectiveness and yet, alleviate the problem of explanation fabrication, FiD-Ex introduces the novel use of sentence markers into pre-trained seq2seq models. Training seq2seq models to decode sentence marker tokens instead of explanation tokens not only guarantees the production of unaltered rationales but also significantly improves explanation metrics on five datasets (Section 7).

Fine-tuning pre-trained models on data-rich intermediate tasks before fine-tuning on classification end tasks has recently been shown to improve end-task performance (Vu et al., 2020; Pruksachatkun et al., 2020), more so in the few-shot setting. We find that this method also extends to seq2seq models, for explanation generation. We fine-tune pre-trained seq2seq models to extract supporting evidence for existing open-domain QA datasets such as Natural Questions (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018), which then improves downstream performance on rationale extraction benchmarks. This approach is motivated by the similarity of the process of gathering supporting facts for QA, to that of rationale extraction for classification tasks. While earlier works on rationale generation (Paranjape et al., 2020; Narang et al., 2020) are limited by the input passage size of pre-trained models and resort to input-passage truncation, FiD-Ex uses the Fusion-in Decoder (FiD) approach (Izacard and Grave, 2020), that separately encodes chunks of long passages and fuses them in the decoder, which further improves performance.

We combine these methods described above to develop FiD-Ex (**Ex**tractive **F**usion-**i**n-**D**ecoder). To summarize, FiD-Ex significantly improves upon the performance and trustworthiness of seq2seq models for rationale generation by 1) reducing their ability to fabricate explanations using sentence markers, 2) extending them to very long input passages, and, 3) intermediate fine-tuning on re-structured existing QA datasets. When applied to the ERASER datasets (DeYoung et al., 2020), a popular benchmark for rationale extraction, FiD-Ex yields significant gains on multiple tasks in terms of explanation metrics: an absolute token-F1 gain of 12.7% on Boolean Question Answering (BoolQ), 33.2% on MovieReviews, 5.3% on Evidence Inference, 2.8% on FEVER, and 2.1% on MultiRC, along with modest gains in terms of task accuracy, over prior work.

2 Related Work

Deep learning models typically function as black boxes offering very little insight into their decision making mechanics. To expose model understanding at various depths, researchers have proposed various structural probing (Tenney et al., 2018; Hewitt and Manning, 2019; Lin et al., 2019) and behavioral probing methods (McCoy et al., 2020; Goldberg, 2019; Warstadt et al., 2019; Ettinger, 2020), as well as input saliency maps to highlight the most important tokens/sentences in the input for each prediction (Serrano and Smith, 2019; Ribeiro et al., 2016; Swanson et al., 2020; Tenney et al., 2019), and input token relationships (Lamm et al., 2020). Alongside, there is work on producing textual rationales (Lei et al., 2016), which are snippets of NL to help explain model predictions. Models may take a pipelined approach, where rationales are first selected as the sole inputs to the prediction stage, either in a supervised (Lehman et al., 2019; Pruthi et al., 2020) or an unsupervised (Paranjape et al., 2020; Bastings et al., 2019; Jain et al., 2020) fashion. Alternatively, rationales can also serve as post-hoc supporting evidence, produced after the model prediction, as a snippet to help users verify the prediction (Yang et al., 2018; Thorne et al., 2018). In this work, we improve upon seq2seq models to produce the latter kind of NL explanations, along with model predictions.

In addition to extractive NL rationales obtained from subsequences of the input text, there is recent work on generating abstractive textual explanations for NLP tasks such as commonsense QA (Rajani et al., 2019) and NLI (Camburu et al., 2018; Kumar and Talukdar, 2020). Latcinnik and Berant (2020) train language models to transparently output their world knowledge as NL tokens, which is then consumed by a light-weight classifier. Narang et al. (2020) use a generative seq2seq T5 model to produce NL explanations token-by-token for the extractive ERASER benchmark, in order to take advantage of multi-task training, i.e., training for task prediction alone, or jointly with explanations if available. Unlike strict input attribution based

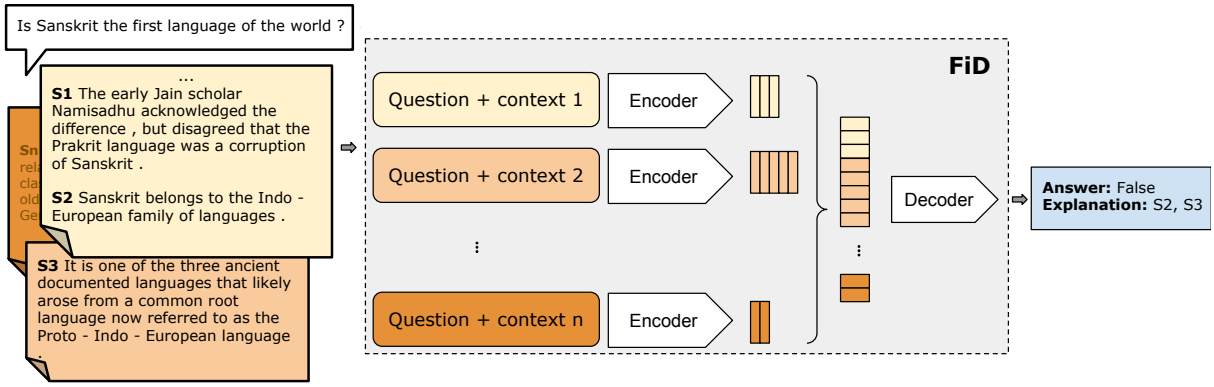

Figure 2: Fusion-in-Decoder architecture for rationale prediction. Each sentence from the passage is marked with sentence markers S1 ... SN. The passage is broken up into C contexts/chunks, which are passed to the encoder. The decoder then attends to the C concatenated and encoded passages to generate the output sequence. The output sequence is the classification token followed by rationale sentence markers.

methods that seldom produce human readable explanations, these models can provide users with more context, keeping with the style of explanation annotations in standard benchmarks such as ERASER. However, such models are susceptible to fabricating explanations to justify even their incorrect predictions, as identified by Camburu et al. (2020) and Wiegreffe et al. (2020). We introduce sentence markers into seq2seq models which alleviates this problem and also significantly improves their rationale extraction performance on sentence-level ERASER benchmark tasks (see Section 4.2).

Multiple prior works (Paranjape et al., 2020; Jain et al., 2020; Narang et al., 2020) have explored methods to improve few-shot rationale generation, to reduce reliance on expensive rationale annotations. We fine-tune FiD-Ex on re-structured intermediate QA datasets to improve its regular and few-shot performance for rationale extraction. Fine-tuning large pre-trained models on intermediate tasks has been shown to be effective by prior work; Phang et al. (2018) use data-rich intermediate NLI tasks to improve target classification tasks; Talmor and Berant (2019) fine-tune on multiple QA datasets to improve the generalizability of QA models. Intermediate fine-tuning (IFT) can also hurt performance (Bingel and Søgaard, 2017). Pruksachatkun et al. (2020) recently present a large-scale study on fine-tuning a pre-trained RoBERTa model on 100 intermediate-target task combinations and use 25 probing tasks to understand the most desirable properties of intermediate tasks and datasets. Vu et al. (2020) explore transferability between 33 NLP tasks and using task embeddings to predict the utility of intermediate tasks, they conclude that intermediate tasks requiring high levels of reasoning and inference abilities are more likely to help, particularly when task data is scarce. Closest to our method is Kung et al. (2020) who use Squad 2.0 as an intermediate task to fine-tune a shared encoder fitted with task-specific classification heads, for the downstream BeerReview and MovieReview rationalization tasks. Our approach is to strategically restructure large open domain QA datasets (Natural Questions and HotpotQA) to make them amenable to IFT of both the encoder and the decoder of pre-trained seq2seq models. This enables the use of exactly the same model architecture for multiple rationale prediction tasks.

3 Modeling

In this section, we develop FiD-Ex, which improves upon seq2seq approaches to jointly produce NL rationales along with model predictions. We illustrate our method using the BoolQ dataset from the ERASER explainability benchmark, which comprises of questions with passages and boolean answers (see Figure 1), together with human annotated rationales (details in Section 4).

Formally, given an input query q and an input passage p comprising sentences $p = \{s_j\}_{j=1}^{N}$, our goal is to produce a prediction y and rationale sentences $\{e_k\}_{k=1}^{K}, e_k \in p, K \ll N$, that justify y.

Narang et al. (2020) fine-tune the pre-trained T5 (Text-to-Text Transfer Transformer) model (Raffel et al., 2019) to auto-regressively produce the prediction and the explanation in a token-by-token fashion. Specifically, their model takes an input of the form "explain {task-name}: q p", represented as a sequence of subword units (Sennrich et al.,

2016) using SentencePiece (Kudo and Richardson, 2018), and is trained to auto-regressively maximize the likelihood of an output sequence represented as "{prediction} explanation: $e_1 \cdots$ explanation: e_K". For example, an input from the BoolQ dataset (Clark et al., 2019) might be represented as "explain boolq: Is Sanskrit the first language of the world? <passage-tokens>", with the output represented as "False explanation: Sanskrit belongs to the Indo-European family of languages. explanation: It is one of the three ..." Such a model can be trained on data, both with and without explanation annotations, by dropping the unavailable parts of the output sequence. This model achieves state-of-the-art explanation performance on several ERASER tasks and serves as a strong baseline which we build upon.

3.1 Sentence Markers

Narang et al. (2020), as well as other works (Camburu et al., 2020), point out that seq2seq models can fabricate reasonable sounding rationales to justify their incorrect predictions. To alleviate this issue, we introduce sentence markers into the input and output to enable the model to learn to generate a rationale sentence as a single unit. This technique has the added benefit that the rationales produced by the model are guaranteed to be strictly extractive at the sentence level, while retaining the performance benefits of a seq2seq architecture. Specifically, we preprocess the input passage p by prefixing each sentence s_i with a sentence marker token `S{i}`. We also train the decoder to output the special sentence marker tokens, instead of NL tokens. Thus, the input is represented as "question: q passage: S1 s_1 S2 $s_2 \cdots$ SN s_N" and the output as "False explanation: $S_{e_1} \cdots$ explanation: S_{e_K}", where S_{e_K} is the marker for e_K. The example from BoolQ would be represented as "explain boolq question: Is Sanskrit the first language of the world passage: S1 <Sent-1> ... SN <Sent-N>" and the output as "False explanation: S2 explanation: S3". Note that these markers are injected as NL text, and would be later split into sub-word units. During inference, sentence markers are produced and mapped back to the corresponding sentences from the input.

3.2 Fusion-in-Decoder Approach

Current approaches typically truncate p to 512 or 1,024 tokens, which is particularly limiting for passages from datasets such as BoolQ, which use very long input passages (> 3000 tokens). To accommodate longer input passages, both for intermediate fine-tuning (see Section 3.3) and target fine-tuning, we use the Fusion-in-Decoder (FiD) architecture of Izacard and Grave (2020) as a replacement for the single encoder-decoder model of Narang et al. (2020). Using FiD, we break p into smaller chunks and encode each chunk independently using the pre-trained T5 encoder (see Figure 2). This expands the effective input length of the encoder, and at the same time, keeps computation resources growing linearly with the number of passages as opposed to quadratically. These separately encoded representations are then fused in the decoder, which then attends to all passage tokens, when producing output tokens. For encoding, we concatenate the query q with each chunk of the input passage p. Further, we also prefix query and context tokens with special tokens, "question:" and "passage:" respectively. Making use of additional context from the passage, without truncation, significantly improves performance on the intermediate fine-tuning tasks as well as on the BoolQ, Movie Reviews and Evidence Inference end tasks (see Table 2). If using sentence markers, they are added to the passage before subdividing into multiple chunks.

Dataset	Train	Val	Test	Toks / Sents
NQ	69,662	4,352	-	1,782 / 66
HotpotQA	180,894	14,810	-	1,649 / 75
BoolQ	6,363	1,491	2,807	3,391 / 165
Movies	1,600	200	200	774 / 37
EVI	7,958	972	959	4,658 / 153
MultiRC	24,029	3,214	4,848	300 / 14
FEVER	97,957	6,122	6,111	288 / 11

Table 1: Dataset split sizes for our intermediate fine-tuning (top) datasets and evaluation (bottom) datasets. We also compare their passage lengths in terms of number of input tokens and sentences.

3.3 Intermediate Fine-tuning (IFT)

Since obtaining rationale annotations for datasets is expensive, we look to fine-tune on existing large datasets to improve target task performance, particularly in the few-shot setting. Specifically, we re-structure open-domain QA (ODQA) datasets with answer span annotations to follow the same input-output structure as our target tasks, i.e., we produce a dataset of (query q, passage p, prediction y, and extractive rationales e) tuples from existing ODQA datasets. The datasets, together with their specific re-structuring methods, are described in Section 4. We present experiments where we

first fine-tune FiD-Ex on a combination of multiple ODQA datasets, and finally, fine-tune on our target evaluation task, in Section 7.

Alternatively, when multiple annotated datasets are available, we can possibly train a universal single model on the combined datasets, that works for all evaluation tasks. We explore this in Section 7.2.

4 Datasets

In this section, we discuss the open-domain QA datasets and our pre-processing steps to prepare them for IFT, as well as, the ERASER rationalizing datasets that we use for evaluation. Table 1 presents the sizes of each dataset split, as well as the average input passage lengths, in terms of the number of tokens and sentences, for both types of datasets.

4.1 Intermediate Fine-Tuning Datasets

Natural Questions (NQ) (Kwiatkowski et al., 2019) comprises real Google search queries with answer-span annotations from Wikipedia pages. Following Lee et al. (2019) we use a subset containing short answers (< 6 tokens). For every question and answer-span annotation, we use the question as q, the segmented Wikipedia passage as p, the answer tokens as the prediction y, and the single sentence containing the answer span as the rationale e. We remove all tables and lists from the Wikipedia passages, but retain section headers.

HotpotQA (Yang et al., 2018) is a multi-hop QA dataset, where each question and answer annotation is accompanied with supporting fact sentence annotations from multiple Wikipedia documents. Similar to NQ, we use the question as q and the answer tokens as the prediction y. Since there are multiple Wikipedia evidence pages, we treat each page as a separate passage p and aggregate the annotated rationale sentences from it as the rationales e. Thus, a single HotpotQA (question, answer) tuple produces as many examples as Wikipedia pages that are part of its supporting facts.

4.2 Evaluation Data

We evaluate on a subset of the datasets from the ERASER benchmark (DeYoung et al., 2020), which comprise an input query and passage, an output class label, and input sentences annotated as rationales. We discuss these datasets in this section.

BoolQ (Clark et al., 2019) comprises questions, whose answer can be either True or False, paired with long Wikipedia passages (> 3,000 tokens), as well as sentence-level rationale annotations (provided by ERASER) that support the answer.

MultiRC (Khashabi et al., 2018) comprises input passages and questions, with multiple-choice answers, with sentence level rationale annotations. It is evaluated as a Boolean QA task by concatenating each answer choice to the question, and assigning a True label to correct choices and False to the rest. All choices use the same set of supporting facts.

MovieReviews (Movies) (Zaidan and Eisner, 2008; Pang and Lee, 2004) contains movie reviews paired with binary positive/negative labels, without a query q (we set it to "What is the sentiment of this review?" in our models). While ERASER provides span-level rationale annotations, we translate these to sentence level annotations following prior work (Paranjape et al., 2020). FiD-Ex can also potentially be trained to output extracted input phrase markers and we leave this to future work.

FEVER (Thorne et al., 2018) The ERASER version of FEVER contains input passages along with claims (q) that must be classified as supported or refuted, based on the passage, together with sentence-level rationale annotations from the input passage.

Evidence Inference (EVI) (Lehman et al., 2019) comprises (intervention, outcome, comparator) triples (concatenated as q) together with randomized controlled trial articles (> 4,000 tokens), with the prediction being whether the intervention significantly increases, decreases, or has no effect on the outcome with respect to the comparator of interest. ERASER provides sentence-level supporting facts on a subset of this dataset.

We do not evaluate on the ERASER datasets of **e-SNLI** and **CoS-E** since they only use single-sentence input passages.

5 Evaluation Metrics

We report Exact Match Accuracy (EM) in terms of exact token match between the predicted class label and the true label, which is equivalent to traditional classification accuracy. To evaluate the explanation quality, we report the following:

Rationale F1 (RF1) is an F1 score over the set of predicted explanation sentences as compared to the set of gold explanation sentences, computing set intersection based on exact sentence match.

Token F1 (TF1) is a token level F1 score between the predicted explanation sentence tokens and the gold explanation sentence tokens, in terms of sets of token positions, by first mapping tokens to token positions in the input passage. This is computed exactly as in Narang et al. (2020), using `spaCy` for tokenization. When using sentence markers, we map the markers back to the original sentences before computing TF1.

Intersection over Union (IOU F1) as described in DeYoung et al. (2020), is computed by first matching up each predicted rationale with a gold rationale, and then computing F1. IOU is similar to RF1, except that it does not use exact match. A prediction and gold sentence match if the size of the overlap of their token positions divided by the size of the union of the token positions is higher than a threshold (we use 0.5). For our models, IOU F1 is very similar in magnitude to RF1.

Other Metrics We do not use human evaluation scores since Narang et al. (2020) found them to be much higher than the automated metrics, and therefore, hard to interpret, in addition to being expensive and noisy. Also, since we aim to provide users with evidence for model predictions, *causal* faithfulness metrics such as comprehensiveness and sufficiency (DeYoung et al., 2020), do not apply.

6 Implementation Details

We use the FiD (Izacard and Grave, 2020) model architecture with T5-base (220M params). We use 1024 input sub-word tokens per context for MultiRC and 512 for the rest. We use a maximum context size of 10 for BoolQ and EVI, and 6 for Movies. We use data distributed training on machines with 8 32-GB GPUs with a batch size of 8 per GPU. We train all models for 20,000 steps using Adam (Kingma and Ba, 2014), with learning rates chosen from $\{1e^{-4}, 1e^{-5}\}$ based on dev performance and use linear decay. We compute dev metrics every 500 steps and select the model with the best TF1 score. We use greedy decoding for the prediction and the explanation. The above settings are used, both for IFT as well as for end-task fine-tuning. For segmenting Wikipedia passages into sentences for NQ, we use Punkt (Kiss and Strunk, 2006) for English from `nltk`. For our evaluation datasets, we used the pre-segmented and pre-tokenized input passages provided by ERASER.

	EM	RF1	IOU F1	TF1
BoolQ				
C=1, No SM	65.2	42.9	46.1	47.0
C=10, No SM	69.4	48.8	51.9	53.3
C=1, With SM	73.6	50.4	50.4	51.1
C=10, With SM	74.6	57.8	57.8	58.3
+ IFT	**76.9**	**59.3**	**59.3**	**59.7**
C=10, With SM, 25%	71.0	51.6	51.6	52.5
+ IFT	**72.9**	**55.1**	**55.1**	**55.7**
Universal	76.3	**57.9**	**57.9**	**58.6**
+ IFT	**77.3**	**57.9**	**57.9**	58.3
Movie Reviews				
C=1, No SM	90.5	19.8	26.5	29.3
C=6, No SM	**98.0**	40.9	51.7	56.6
C=1, With SM	89.0	55.5	55.8	57.5
C=6, With SM	97.5	**64.3**	**64.3**	**65.9**
+ IFT	97.0	64.0	64.1	65.5
C=6, With SM, 25%	**97.0**	61.0	61.1	62.3
+ IFT	96.5	**61.5**	**61.6**	**63.2**
Universal	97.0	**64.6**	**64.6**	**66.6**
+ IFT	**98.0**	**64.6**	**64.6**	66.5
Evidence Inference				
C=1, No SM	66.3	14.7	15.1	14.6
C=10, No SM	**75.8**	27.0	27.4	27.1
C=1, With SM	63.1	29.8	29.8	29.8
C=10, With SM	75.2	50.7	50.7	50.9
+ IFT	74.7	**52.0**	**52.1**	**52.1**
C=10, With SM, 25%	**73.0**	46.3	46.4	46.4
+ IFT	70.6	**47.3**	**47.3**	**47.6**
Universal	75.3	50.2	50.4	50.1
+ IFT	**77.4**	**51.4**	**51.5**	**51.4**
MultiRC				
C=1, No SM	78.1	67.1	68.0	67.8
C=1, With SM	78.5	72.2	72.2	71.9
+ IFT	**79.8**	**72.4**	**72.4**	**72.0**
C=1, With SM, 2k	76.4	**70.2**	**70.2**	**69.8**
+ IFT	**76.9**	69.5	69.5	69.2
Universal	80.0	**72.6**	**72.6**	72.1
+ IFT	**80.6**	72.4	72.4	**72.2**
FEVER				
C=1, No SM	92.9	69.8	70.9	70.7
C=1, With SM	92.9	83.5	83.5	83.4
+ IFT	**93.1**	**84.1**	**84.1**	**84.0**
C=1, With SM, 2k	**88.6**	80.9	80.9	80.7
+ IFT	88.2	**81.4**	**81.4**	**81.2**
Universal	94.1	87.9	87.9	87.8
+ IFT	**94.4**	**88.2**	**88.2**	**88.0**

Table 2: Performance of FiD-Ex using sentence markers (SM), larger contexts (C), and intermediate fine-tuning (IFT) on 5 ERASER tasks in the fully supervised and low-resource settings, alongwith that of a single universal model (trained on all datasets combined).

7 Results and Discussion

We compare the performance of different variants of our FiD-Ex model using all evaluation metrics on five ERASER datasets, in Table 2. The

Model	BoolQ			Movie Reviews			Evidence Inference		
	EM	IOU F1	Token F1	EM	IOU F1	Token F1	EM	IOU F1	Token F1
Bert-to-Bert	54.4	5.2	13.4	86.0	7.5	14.5	70.8	45.5	46.8
WT5 Base	–	–	–	**98.0**	–	32.7	–	–	–
IB Supervised	63.4	32.3	19.2	85.4	43.4	28.2	46.7	13.3	10.8
FiD-Ex Base	**76.9**	**59.3**	**59.7**	97.5	**64.3**	**65.9**	**74.7**	**52.1**	**52.1**

Model	MultiRC			FEVER		
	EM	IOU F1	Token F1	EM	IOU F1	Token F1
Bert-to-Bert	63.3	41.6	41.2	87.7	83.5	81.2
WT5-Base	77.8	–	69.9	–	–	–
IB Supervised	66.4	54.4	54.0	88.8	66.6	63.9
FiD-Ex	**79.8**	**72.4**	**72.0**	**93.1**	**84.1**	**84.0**

Table 3: Performance of our best FiD-Ex model (multi-context input, sentence markers, and IFT) compared with prior work. For WT5, we use their base model since we report all our metrics using T5-base. IOU F1 for IB is reported using a threshold of 0.1 whereas we report all our IOU metrics using a stricter threshold of 0.5.

first row for each dataset can be viewed as our re-implementation of Narang et al. (2020) i.e., T5 with a single context, without sentence markers. We use Token-F1 (TF1) to describe all results, but observe similar trends for the other rationale metrics. We report all gains in *absolute* percentage points.

Sentence Markers The addition of sentence markers leads to a large improvement in explanation metrics on all datasets as compared to generating raw tokens, demonstrating the capabilities of pre-trained seq2seq models to select *scattered input markers*; For the single context case, BoolQ TF1 improves by 4.1%, Movies by 28.2%, EVI by 15.2%, MultiRC by 4.1% and FEVER by 12.7%. Additionally, it also provides the desirable guarantee of being extractive by eliminating the problem of fabricated rationales that seq2seq models are susceptible to (Appendix B presents examples of fabrication). Furthermore, while WT5 (Narang et al., 2020) yielded a TF1 of 0 on MultiRC when trained on less than 10,000 examples, FiD-Ex obtains a TF1 of 69.8% with just 2,000 examples owing to the use of sentence markers.

Increased Passage Size Using FiD's multiple context encoders instead of the input truncation methods of prior work, helps significantly improve performance. When also using sentence markers, BoolQ TF1 improves by 7.2%, Movies by 8.4% and EVI by 21.1%. This is accompanied by task EM gains of 8.5% in Movies and 12.1% in EVI. Input passages in MultiRC and FEVER are not long enough to benefit significantly from increased passage size. The gains from increasing passage size are orthogonal to the gains by sentence markers, i.e., explanation metrics improve with additional context with or without using sentence markers (Table 2). Similarly, sentence markers improve performance for both single and multi-contexts.

Intermediate Fine-tuning (IFT) and Few-shot Performance We perform IFT using sentence markers on a combined dataset of NQ and HotpotQA, re-formatted for rationale extraction tasks. Final fine-tuning on the full training sets of our evaluation tasks improves TF1 by 1.4% for BoolQ and 1.2% for EVI. To evaluate IFT in the few-shot setting, we fine-tune using 25% data for the BoolQ, Movies, and EVI tasks following Paranjape et al. (2020) and 2,000 examples for tasks with bigger datasets, viz., MultiRC and FEVER. We see an improvement of 3.2% TF1 on BoolQ and 1.2% on EVI. This is desirable since obtaining labeled rationale annotations is expensive. We do not observe any performance improvement for Movies, MultiRC, and FEVER with IFT. While our few-shot experiments used 25% data to compare with prior work, IFT may show more marked improvements with just 10-100 examples. While IFT on NQ or HotpotQA alone improves performance, we find that combining the datasets yields best results.

7.1 Comparison with Prior Work

In Table 3 we compare our best fully supervised model for each dataset, with prior works that share the best performance on ERASER tasks:

Bert-to-Bert (B2B) is the supervised pipeline of DeYoung et al. (2020) that comprises an independently trained rationale extractor, and an answer prediction model on the extracted rationales.

Error Type	% Cases
Overlap and Adequate	36
Overlap and Inadequate	4
Over-Prediction	30
No-overlap and Inadequate	12
No-overlap and Adequate	8
Prediction not in input	4
Input Truncated	6

Table 4: Distribution of error types in 50 randomly sampled examples with a non-perfect RF1 score, from the dev set of BoolQ, using our best FiD-Ex model.

The Information Bottleneck (IB) approach of Paranjape et al. (2020), which jointly trains an explainer that predicts sparse binary masks over input sentences, and a prediction model on the residual sentences. Although they only report supervised results using 25% training data, their model achieves similar performance even with 100% training data.

WT5-Base is the base version of the seq2seq model of Narang et al. (2020).

Overall, we outperform prior work on explanation metrics (using TF1) on BoolQ (+40.5% from IB), Movies (+33.2% from WT5), EVI (+5.3% from B2B), MultiRC (+2.1% from WT5), and FEVER (+2.8% from B2B). We also improve Task Accuracy on BoolQ (+13.5% from IB), EVI (+3.9% from B2B), MultiRC (+2.0% from WT5), and FEVER (+4.3% from IB). In summary, FiD-Ex significantly improves the state-of-the-art on five ERASER datasets, in fully supervised and few-shot settings, with each component from Section 3 individually contributing to overall performance.

7.2 Universal Model

With the goal of deploying one single model that can perform all 5 ERASER tasks, we train a model on their combined training sets, with SM and $C = 10$, and evaluate on each test set (see Table 1). Each training example is prefixed with a token denoting the dataset that it came from as described in Section 3. Despite the lack of individual fine-tuning, this universal model outperforms the best fine-tuned models by 4% on FEVER and is within $\pm 1\%$ of the best model performance on the other datasets. Training on a large combined dataset of related tasks, when available, reduces reliance on IFT to improve performance (which primarily benefits only EVI in this scenario). Overall, this result highlights a key advantage of the seq2seq format, that naturally enables effective data sharing among multiple related tasks (Raffel et al., 2019).

7.3 Error Analysis

We conduct an error analysis on predictions from our best FiD-Ex model on 50 random examples from the valid set of BoolQ, which have non-perfect RF1 score. (Table 4). The two largest error types are: 1) *Overlap and Adequate* (36%): the set of predicted explanations is adequate by itself and overlaps with the true explanations, i.e., the true explanation set contains redundancies, and 2) *Over-prediction* (30%): the set of predictions is a strict superset of the true explanations. Other sources of errors are *Overlap and Inadequate* (4%) - when the predictions are inadequate but overlap with the true explanations, *No-overlap and Adequate/Inadequate* - when the predictions have no overlap with the true explanations and are either still adequate (8%) or inadequate (12%). Since ERASER provides only one of the multiple possible explanation sets, 8% non-overlapping predictions happen to be adequate. *Prediction not in input* (4%) - when sentence markers that do not exist in the input are predicted, and *Input Truncated* (6%) - when the true explanation sentences are truncated out of the model input, which still happens for very long inputs even with a context size of 10. We present illustrative examples of these error cases in the Appendix. Promising focus areas for future work include addressing model tendencies for over-prediction (30% of cases) and inadequate non-overlapping predictions (12% of cases).

8 Conclusion

In this paper, we develop general methods to improve the performance of large pre-trained seq2seq models for jointly producing NL rationales and answer predictions. Specifically, we introduce sentence markers into seq2seq models to tackle explanation fabrication, we enable larger input passage sizes using the Fusion-in-Decoder architecture, and we infuse knowledge by fine-tuning on restructured QA datasets. We show that a universal model can perform favourably compared to the best task-specific fine-tuned models. Our methods improve the state of the art on rationale extraction metrics and task accuracy on multiple ERASER benchmarks while reducing the extent to which seq2seq models fabricate explanations to justify incorrect predictions, thereby improving the reliability and verifiability of the generated rationales.

References

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics. Released under the CC-BY-SA 3.0 license (https://creativecommons.org/licenses/by-sa/3.0/).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. Nile: Natural language inference with faithful natural language explanations. *arXiv preprint arXiv:2005.12116*.

Po-Nien Kung, Tse-Hsuan Yang, Yi-Cheng Chen, Sheng-Siang Yin, and Yun-Nung Chen. 2020. Zero-shot rationalization by multi-task transfer learning from question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2187–2197.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Zachary C Lipton. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 271–es, USA. Association for Computational Linguistics.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Weakly-and semi-supervised evidence extraction. *arXiv preprint arXiv:2011.01459*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. *arXiv preprint arXiv:2005.13111*.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2018. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasovic, and Noah A Smith. 2020. Measuring association between labels and free-text rationales. *arXiv preprint arXiv:2010.12762*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

A Error Analysis on BoolQ

We present examples for each error type from our error analysis of model predictions using 50 randomly chosen examples from the dev set of BoolQ to illustrate the cases with a non-perfect Rationale F1 score. We have preserved the sentence markers (SM) in the document to help locate the gold and predicted sentences easily The error types are:

1. *Overlap and Adequate* - Predicted explanations are adequate and overlap with the true explanations

2. *Overlap and Inadequate* - Predicted explanations are inadequate but overlap with the true explanations

3. *Over-prediction* - Predicted explanations are a strict superset of the true explanations

4. *No overlap and Inadequate* - Predicted explanations are inadequate and do not overlap with the true explanations

5. *No overlap and Adequate* - Predicted explanations are adequate but do not overlap with the true explanations

6. *Prediction not in input* - Predicted explanation sentence markers are not in the input

7. *Input Truncated* - True explanation sentence markers are not in input

Legend: **Sentence Marker (SM)**, Correctly Predicted SM, Missed SM, Over-predicted SM

Overlap and Adequate

Question: is a woodchuck and a groundhog the same
Gold Answer: True
Predicted Answer: True
Gold Rationales: ['S1', 'S2', 'S3', 'S4', 'S5', 'S6', 'S7', 'S8', 'S9']
Predicted Rationales': ['S0', 'S1', 'S2', 'S3']
Document: **S0** GROUNDHOG **S1** The groundhog (Marmota monax) , also known as a woodchuck , is a rodent of the family Sciuridae , belonging to the group of large ground squirrels known as marmots . **S2** It was first scientifically described by Carl Linnaeus in 1758 . **S3** The groundhog is also referred to as a chuck , wood - shock , groundpig , whistlepig , whistler , thickwood badger , Canada marmot , monax , moonack , weenusk , red monk and , among French Canadians in eastern Canada , siffleux . **S4** The name " thickwood badger " was given in the Northwest to distinguish the animal from the prairie badger . **S5** Monax (Móonack) is an Algonquian name of the woodchuck , which meant " digger " (cf . **S6** Lenape monachgeu) . **S7** Young groundhogs may be called chucklings . **S8** Other marmots , such as the yellow - bellied and hoary marmots , live in rocky and mountainous areas , but the groundhog is a lowland creature . **S9** It is found through much of the eastern United States across Canada and into Alaska DESCRIPTION Section::::Description . **...** **S159** * Woodchuck (Groundhog) , Missouri Conservation Commission * Breeding and Experimental Facility for Woodchucks

Overlap and Inadequate

Question: are all mass air flow sensors the same
Gold Answer: False
Predicted Answer: False
Gold Rationales: ['S4', 'S5', 'S6', 'S7', 'S8']
Predicted Rationales': ['S0', 'S1', 'S2', 'S3', 'S4']
Document: 'S0 MASS FLOW SENSOR S1 A mass (air) flow sensor (MAF) is a sensor used to determine the mass flow rate of air entering a fuel - injected internal combustion engine . S2 The air mass information is necessary for the engine control unit (ECU) to balance and deliver the correct fuel mass to the engine . S3 Air changes its density with temperature and pressure . S4 In automotive applications , air density varies with the ambient temperature , altitude and the use of forced induction , which means that mass flow sensors are more appropriate than volumetric flow sensors for determining the quantity of intake air in each cylinder . S5 There are two common types of mass airflow sensors in use on automotive engines . S6 These are the vane meter and the hot wire . S7 Neither design employs technology that measures air mass directly . S8 However , with additional sensors and inputs , an engine ś ECU can determine the mass flow rate of intake air . **...** **S103** REFERENCES EXTERNAL LINKS * A Hot Film sensor with theory of operation * A video example of cleaning a MAF sensor * An example of how to clean a MAF sensor , **S104** 3 wire **S105** * How To Test a MAF

Over − Prediction

Question: was kentucky a southern state in the civil war
Gold Answer: False
Predicted Answer: False
Gold Rationales: ['S4', 'S5', 'S6', 'S7']
Predicted Rationales': ['S0', 'S1', 'S2', 'S3', 'S4', 'S5', 'S6']
Document: S0 KENTUCKY IN THE AMERICAN CIVIL WAR Kentucky was a border state of key importance in the American Civil War . S1 President Abraham Lincoln recognized the importance of the Commonwealth when , in a September 1861 letter to Orville Browning , he wrote : I think to lose Kentucky is nearly the same as to lose the whole game . S2 Kentucky gone , we can not hold Missouri , nor Maryland . S3 These all against us , and the job on our hands is too large for us . S4 We would as well consent to separation at once , including the surrender of this capitol . S5 Kentucky , being a border state , was among the chief places where the " Brother against brother " scenario was prevalent . S6 Kentucky officially declared its neutrality at the beginning of the war , but after a failed attempt by Confederate General Leonidas Polk to take the state of Kentucky for the Confederacy , the legislature petitioned the Union Army for assistance . S7 After early 1862 Kentucky came largely under Union control . **...** **S128** Union ironclads routed the Confederate river gunboats on the Mississippi River during the Battle of Lucas Bend on January 11 , forcing them back to Columbus .

No overlap and Adequate

Question: is row row row your boat a masonic poem
Gold Answer: False
Predicted Answer: False
Gold Rationales: ['S0', 'S1', 'S2', 'S3']
Predicted Rationales': ['S33', 'S34', 'S35', 'S36', 'S37']
Document: **S0** ROW , ROW , ROW **S1** YOUR BOAT " Row , Row , Row Your Boat " is an English language nursery rhyme and a popular children 's song . **S2** It can also be an " action " nursery rhyme , whose singers sit opposite one another and " row " forwards and backwards with joined hands . **S3** It has a Roud Folk Song Index number of 19236 . **...** **S33** ly , ORIGINS Section::::Origins . **S34** It has been suggested that the song may have originally arisen out of American minstrelsy . **S35** The earliest printing of the song is from 1852 , when the lyrics were published with similar lyrics to those used today , but with a very different tune . **S36** It was reprinted again two years later with the same lyrics and another tune . **S37** The modern tune was first recorded with the lyrics in 1881 , mentioning Eliphalet Oram Lyte in The Franklin Square Song Collection but not making it clear whether he was the composer or adapter . **...** **S42** Don Music , a muppet character in Sesame Street , changed the lyrics to feature a car instead of a boat . **S43** Versions include : And : NOTES AND REFERENCES

No overlap and Inadequate

Question: are there mountains in the state of indiana
Gold Answer: False
Predicted Answer: True
Gold Rationales: ['S107', 'S108']
Predicted Rationales': ['S84', 'S85', 'S86']
Document: **S0** GEOGRAPHY OF INDIANA **S1** The geography of Indiana comprises the physical features of the land and relative location of U.S. State of Indiana . **...** **S84** Rural areas in the central portion of the state are typically composed of a patchwork of fields and forested areas . **S85** The geography of Central Indiana consists of gently rolling hills and sandstone ravines carved out by the retreating glaciers . explain boolq question: are there mountains in the state of indiana passage: **S86** Many of these ravines can be found in west - central Indiana , specifically along Sugar Creek in Turkey Run State Park and Shades State Park . **...** **S107** PHYSIOGRAPHY Section::::Physiography . **S108** Indiana is broken up into three main physical regions : The Great Lakes Plain in the northern third of the state , the Tipton Till Plain in the central third , and the Southern Hills and Lowlands region in the southern third . **...** **S136** * Midwestern United States NOTES REFERENCES

Prediction not in Input

Question: is costa rica part of the ring of fire
Gold Answer: True
Predicted Answer: True
Gold Rationales: ['S121', 'S122', 'S123', 'S124', 'S125', 'S126', 'S127', 'S128']
Predicted Rationales': ['S261', 'S262', 'S263', 'S264', 'S265', 'S266', 'S267', 'S268', 'S269', 'S270']
Document: **S0** RING OF FIRE **S1** The Ring of Fire is a major area in the basin of the Pacific Ocean where many earthquakes and volcanic eruptions occur . **...** **S121** AMERICA COSTA RICA Section::::Central America . **S122** Section::::Costa Rica . **S123** The Volcanological and Seismological Observatory of Costa Rica (OVSICORI) at the National University of Costa Rica , in Spanish Observatorio Vulcanológico y Sismológico de Costa Rica (OVSICORI) have a dedicated team in charge of researching and monitoring the volcanoes , earthquakes , and other tectonic processes in the Central America Volcanic Arc . explain boolq question: is costa rica part of the ring of fire passage: **S124** In 1984 , the OVSICORI - A initiated the operation of a seismographic network designed to monitor seismic and volcanic activity throughout the national territory . **S125** Currently , the seismographic network has an analog and a digital registration system . **S126** The latter enables online analysis of seismic signals , allowing to expedite the analysis of signals and the study using modern computerized methods . **S127** Poás Volcano is an active stratovolcano located in central Costa Rica ; it has erupted 39 times since 1828 . **S128** On February 25 , 2014 , a webcam from the OVSICORI captured the moment a dark cloud exploded about in the air from a massive crater of the Poás Volcano . **...** **S138** A few other active volcanoes in northern Mexico are related to extensional tectonics of the Basin and Range Province , which splits the Baja California peninsula from the mainland .

Input Truncated

Question: did the harry potter movies win any oscars
Gold Answer: False
Predicted Answer: True
Gold Rationales: ['S297', 'S298', 'S299', 'S300']
Predicted Rationales': ['S261', 'S262', 'S263', 'S264', 'S265']
Document: **S0** HARRY POTTER (FILM SERIES) **S1** Harry Potter is a British - American film series based on the Harry Potter novels by author J. K. Rowling . **S2** The series is distributed by Warner Bros. and consists of eight fantasy films , beginning with Harry Potter and the Philosopher 's Stone (2001) and culminating with Harry Potter and the Deathly Hallows – Part 2 (2011) . **...** **S125** CAST AND CREW Section::::Cast and crew . **S126** Aside from the three lead actors , other notable cast members include Robbie Coltrane as Rubeus Hagrid , Tom Felton as Draco Malfoy , Alan Rickman as Severus Snape , and Dame Maggie Smith as Minerva McGonagall .

B Extent of Explanation Fabrication

We measure the fraction of explanations produced by the WT5 model (Narang et al., 2020) (C=1) that do not exactly match to any input sentence, for each dataset in Table 5. These are an indication of fabricated sentences in the output of WT5. In contrast, FiD-Ex, which produces sentence markers, never fabricates sentences.

ERASER dataset	% Sentences
BoolQ	8%
MultiRC	3.9%
Movies	60%
FEVER	21%
Evidence Inference	69%

Table 5: Fraction of explanation sentences produced by the WT5 model (Narang et al., 2020) that do not exactly match to any input sentence (with C=1), on five ERASER datasets.

In Table 6, we present an example of an explanation fabricated by WT5. The sentence in the output of WT5 is fabricated, and an alternate sentence with the same beginning exists in the input passage.

Input Movie Review:
note : some may consider portions of the following text to be spoilers . be forewarned . " all the world 's a stage and all the men and women merely players they have their exits and their entrances and one man in his time plays many parts " - excerpt from as you like it , act ii , scene 7 when william shakespeare penned this passage , he could not have possibly envisioned a world in which the domestic activites in an abode would be broadcast across the continent , or where women would install webcams in their apartments in order to convert voyeurism into cash . ... for the first time in his life , it is beginning to dawn on truman that things are not what they appear to be . ...
Output explanations by WT5
... for the first time in his life, truman shows some of the oddest, most upbeat and affectionate attitudes towards women and men in general ....

Table 6: An example of explanation fabrication by the WT5 model (Narang et al., 2020) on an example from the MovieReviews dataset.